IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-11 E-ISSN: 2347-2693

Overview on Data Mining in Social Media

C.Amali Pushpam[1] , J.Gnana Jayanthi[2*]

[1]Dept. of Computer Science, Rajah Serfoji College, Tamil Nadu, India
[2*]Dept. of Computer Science, Rajah Serfoji College, Tamil Nadu, India

Corresponding Author: joemarycap@gmail.com, Tel.: +91 9600627074

Available online at: www.ijcseonline.org

Received: 10/Oct/2017, Revised: 27/Oct/2017, Accepted: 19/Nov/2017, Published: 30/Nov/2017

Abstract - Knowledge plays a vital role in every sphere of human life. Data Mining supports to acquire knowledge by discovering pattern / correlations among data. This information is applied in various applications like business, education, social media, medical, Agriculture etc. Data mining field has attained enormous success from its inception to the present level. Also it faces many issues especially while handling social media data. Social media is one of the important sources that provide huge volume of data that are unstructured and heterogeneous. Handling this data is really a very big challenge to the researchers. At present, a number of data mining algorithms and techniques are available with their own merits and demerits. Finding a suitable algorithm for a particular application is a very big challenge. This paper imparts many issues in data mining and also focuses scope of the data mining in social media which will be helpful in the further research.

Keywords - Data mining, social media, clustering, classification

I. INTRODUCTION

Due to rapid development in Information and Communication Technology, the amount of data available to users has been increasing exponentially from the range of Tera to Exa bytes. The year 2017 turned out to be a strong year for social media and apps companies, with a large number of social media users reaching 2.34 billion, according to statista.com. Due to apps and social media users, huge volumes of data are available in different format like text, video, audio, images, graphics, etc. These data are stored in different types of repository. Because of this data throng the humanity is suffering from new syndrome called "Data Rich and Information Poor". To make use of this resource more effective, retrieval of data is not only enough. It requires a special tool to summarize, analyze, extract the information and discover the pattern and correlation among data which are the real challenges faced by researchers. The solution to above all is "Data Mining". Data mining refers to "Mining of Information" ie. pattern / rule from data that could not be found by manual analysis alone, by applying statistical analysis, artificial intelligence, and machine learning technologies.

In this paper, we have analyzed several data mining techniques particularly the three most applied techniques i.e SVM, BN and DT in the area of social media. Among these three, SVM is the most frequently used algorithm. Its merits and demerits are analyzed.

This paper is drafted to present with seven sections briefing the evolution of data mining, parameters and process, architecture and the types of data mining system in section II, concising the various algorithms proposed for data mining, by pointing out merits and demerits of each of them in section III, reviewing the literature study in section IV, summarizing real time applications of data mining in section V, highlighting the challenges and issues in section VI and future research work in section VII.

II. DATA MINING

Data Mining is referred to as Information Harvesting / Knowledge Mining / Knowledge Discovery in Databases / Data Dredging / Data Pattern Processing / Data Archaeology / Database Mining, Knowledge Extraction and Siftware. Data Mining is a process of analyzing data from many different dimensions or angles and summarizing it into useful information that can be applied in different fields to take proper decision. It increases profits and cuts costs, or both. Technically, data mining is the computing process of discovering patterns or correlations in large relational databases involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [1].

A. Evolution

Bayes' theorem published in the year 1763, is fundamental to data mining and probability. It allows understanding of complex realities based on estimated probabilities. Regression

analysis introduced in the year 1805 is used to estimate the relationship among variables. Regression is one of the key tools in data mining. In 1936, Computer age started where collection and processing of large amounts of data were possible. The basis for object oriented concepts started in the early 1960s. From that period onwards, Programmers started to give importance to data rather than coding. Data plays an important role in all the fields and started to attract the attention of researchers. Due to fast growth of technology, huge volumes of data are available. As a result researchers put their interest in managing and analyzing this data. So the process of extraction of pattern from data has occurred for centuries. 1970s makes possible to store and query tera bytes and peta bytes of data with sophisticated database management systems. The term "Knowledge Discovery in Databases" (KDD) was found in the year 1989. In 1990s, the "data mining" concept appeared in the database. Retail companies and the financial communities apply data mining to analyze data and recognize trends to identify customers' behavior, predict stock prices and increase their customer base. As size and complexity of data set increases rapidly, ordinary tools for data analysis is not enough. Hence data analysis has been enhanced with indirect, automated data processing, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Though there are more number of data mining algorithms under various techniques [2], still research is going on in this area.

B. Parameters

Data mining parameters include sequence or path analysis, classification, clustering and forecasting. Sequence or path analysis parameters look for patterns where one event leads to another later event. A sequence is an ordered list of sets of items, a common data type used in data bases.

A Classification parameter looks for new patterns which result in a change in the way the data is organized. Classification algorithms predict variables based on other factors within the database. Clustering parameters document groups of facts that were previously unknown. Clustering groups a set of objects and aggregates them based on similarities existing in objects. Fostering parameters discover patterns in data that predict about the future, also known as predictive analysis [3].

C. Processes

The processes involved in the knowledge discovery are (i) Data cleaning (to remove noise or irrelevant data), (ii) Data integration (where multiple data sources may be integrated / combined), (iii) Data selection (where data relevant to user's request / analysis task are retrieved from the database), (iv) Data transformation (where data are transformed or consolidated into suitable forms for mining, by performing summary or aggregation operations, for instance), (v) Pattern discovery (Patterns are discovered by applying intelligent methods), (vi) Pattern evaluation (identify the truly interesting patterns representing knowledge based on some interestingness measures), (vii) Knowledge presentation (visualize the patterns in different forms) [4].

D. Architecture

The architecture of a data mining system has the following major components [5] as depicted in figure Fig:1

Database / Data Warehouse / Information Repository:

This is a store house where huge volumes of data are available for data mining for eg: a set of databases, data warehouses, spread sheets or other kinds of information repositories. In the processes of data mining , the first two processes namely data cleaning and data integration are performed on the data.

Database / Data Warehouse Server:

The database / data warehouse server fetch the relevant data, based on the user's request through the data mining process called selection and transformation.

Knowledge Base:

This is the domain knowledge that includes concept hierarchies, user beliefs, thresholds, and metadata that is used to guide the search, or evaluate the interestingness of resulting patterns.

Data Mining Engine:

This is essential in the data mining system to discover pattern by using a set of functional modules for tasks such as characterization, association analysis, classification, evolution and deviation analysis.

Pattern Evaluation Module:

This component is integrated with the mining module and employs interestingness thresholds stored in the knowledge base so as to focus the search towards interesting patterns. Pattern evaluation module plays an important role in efficient data mining to confine the search to only the interesting patterns.

Graphical User Interface:

User can communicate with data mining system through this module and interact with system by providing information through query or task to help focus the search. In addition, this component allows the user to browse database, visualize the patterns in different forms and evaluate mined patterns.

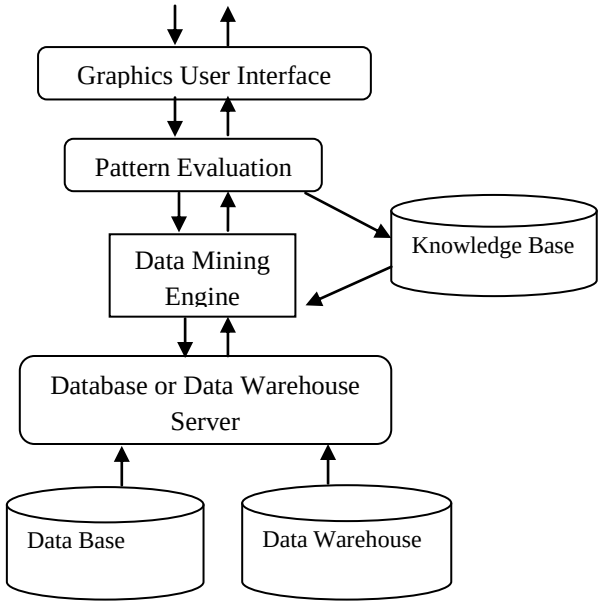

Figure 1 Architecture of Data Mining System

Based on the (i) database used, (ii) kind of information to be extracted and (iii) type of techniques and tools applied, Data Mining System are categorized [6] as listed below.

Type of data source mined

In an organization a huge amount of data are available in different data repositories and in different formats (audio/video, text format etc) and therefore the data are to be classified according to their type.

Data model

Presently different types of data mining models such as Relational data model, Object model, Object Oriented data model, Hierarchical data model/W data model, etc., are available. The data mining system also classifies the data, from the models.

Kind of knowledge discovered

This classification based on data mining functionalities, such as association, clustering, classification, characterization, discrimination, etc. Some systems provide hybrid data mining functionalities together.

Mining techniques used

Based on data analysis approach such as machine learning, neural networks, genetic algorithms, etc classification is done .

III - REAL TIME APPLICATION AREAS

As the importance of data mining continues to grow, it has been used in many applications sectors like sales / marketing, banking, insurance, telecommunication, fraud detection, finance, education sector, medical and so on [6], [7], [8], [9], [10]. Some of the noteworthy applications sectors are listed below.

- *Data Mining in Education Sector*

Educational Data Mining (EDM) is an emerging interdisciplinary research field. Educational Data Mining refers to techniques and tools applied on information generated from educational settings related to students' learning activities and investigate scientific questions within educational research. It is very much useful to understand students' learning behaviors and the settings which they learn in. Use student's data to analyze their learning behavior to predict the results.

- *Data mining in Medicine*

An Enormous Electronic Health Records (EHRs) are available in medical field. Accuracy is considered an extremely important factor while handling these EHRs as it is related with patient's health. DM can generate information that can be useful to all stakeholders in health care, including patients by identifying effective treatments and best practices.

- *Data Mining in Agriculture*

Agriculture is the core of human life. Recently a lot of problems are arising in this field. Applying data mining in agriculture is a novel research field. The idea of using patterns embedded in huge volume of data solving complex agriculture problems and predicting future trends of agricultural processes. For example soil water parameters for a certain soil type can be estimated knowing the behavior of similar soil types.

- *Data mining in Marketing*

Due to rapid fluctuation in the value of currency and dynamic customers' behavior, it is very difficult to take investment decision in business. Also stock market is also being generating huge volume of data. This nature has attracted researchers to apply mining on these data and find patterns to predict the probability of purchasing the product and the future trend of business.

- *Data mining in Social Media*

Social media mining analyzes and extracts patterns or correlations or trends from raw social media data e.g., social media usage, online behaviors, sharing of content, connections between individuals, online buying behavior, etc. These patterns give valid information to companies, governments and non - profit organizations, to design their strategies or introduce new programs / products / services.

IV. ALGORITHMS

According to Nikita Jain, et. al (2013), data mining is collecting relevant information from unstructured data. Hence it helps to achieve specific objectives. The purpose of data mining effort is normally either to create a descriptive model or a predictive model. A descriptive model presents, in a concised form, the main characteristics of the data set. The purpose of a predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable;

the target variable [11]. Data mining model [12] is shown in figure fig.2.

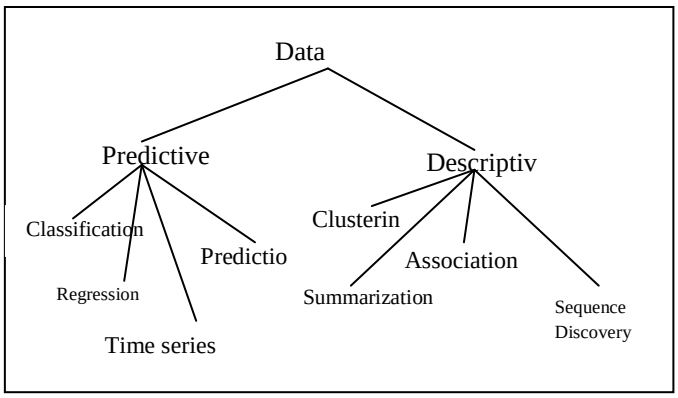

Figure 2 Data Mining Model

❖ *Association*

In Association, a simple correlation between two or more items, often of the same type is used to identify patterns. It is very familiar and better known straight forward data mining technique.

❖ *Classification*

It is a process in which classification is done based on the supervised learning (i.e. desired output for a given input is known) on the training set and values (class label) using a decision tree, neural network and classification rule (IF-Then). For example the classification rule can be applied on the past records of the students who left from university and evaluate them. Using these techniques, the performance of the students can be easily identified.

❖ *Prediction*

It is one of the data mining techniques to analyze past events or instances and predict the future of an event like to predict the failure of components or machineries or to identify fraud or to predict company profits. It discovers the relationship between independent variables and the relationship between dependent and independent variables [3].

❖ *Clustering*

It is a way of finding similarities between data according to their characteristics and form the cluster based on the unsupervised learning (i.e. desired output for a given input is not known). Image processing, pattern recognition, city planning etc., are some of the examples under clustering

❖ *Sequential Patterns*

It is a data mining technique used to identify trends or regular occurrences of similar events. For example, with customer data, customer's behavior can be identified [3].

❖ *Time Series Analysis*

It is a method of using statistical techniques to explain a time-dependent series of data points and generate predictions (forecasts) for future events based on known past events such as in stock market [12], [13].

❖ *Regression*

It is used to map a data item to a real valued prediction variable [14]. In the regression techniques target values are known. For example, one can predict the child behavior based on family history.

❖ *Summarization*

Summarization is abstraction of data, i.e., a set of relevant tasks and gives an overview of data. For example, long distance race can be summarized total minutes, seconds and height.

An algorithm in data mining or machine learning creates a model by analyzing input data and extracting specific types of patterns or correlations. This output is analyzed over many iterations to find the optimal parameters for creating model. These parameters are then applied across the entire data sets to extract actionable patterns. Choosing the best algorithm for a specific analytical task can be a challenge. Different algorithms can be used to perform a same task, each algorithm produces different results, and some algorithms can produce more than one type of result. More number of algorithms are available in the field to fulfill the needs of users. Some of the top most algorithms are: 1) Support vector machines (SVM), 2) Bayesian Networks (BN), 3) Decision Tree (DT), 4) C4.5, 5) K-Nearest Neighbors (KNN) 6) k-means, 7) Apriori, 8) Expectation-Maximization (EM), 9) PageRank, 10) AdaBoost, 11) Naive Bayes, 12) CART

Below mentioned issues are frequently observed in social media data. Data velocity or incoming of data is enormous and dynamic nature of data is also unpredictable. The massive data available in social media is unstructured and heavily used by youngsters. So it is a heavy duty imposed on researchers to find out suitable algorithms to fetch correct patterns / trends / correlations existing among data which are helpful to youngsters for their bright future as well as the nation. 19 data mining techniques had been applied by researchers in the area of social media. Among them SVM, BN, and DT are the most applied techniques in the area of social media.

1. Support Vector Machines (SVM)

Support vector machine (SVM) classify data into 2 classes using the concept of a hyper plane. SVM is similar to C4.5 except SVM doesn't use decision trees at all. For a simple classification task with just 2 features, the hyper plane can be a line,

$$y = m x + b.$$

For example: Take a bunch of balls of red and blue colors on a table which aren't too mixed together. In this case they can be separated with the stick. When a new ball is added on the table, by knowing which side of the stick the ball is on, its color can be predicted. Here the balls represent data points, and the red and blue colors represent 2 classes. The stick represents the simplest hyper plane which is a line. This is a supervised learning, since a dataset is used to first teach the

SVM about the classes. Only then is the SVM capable of classifying new data. SVM along with C4.5 are generally the classifier. No classifier will be the best in all cases. Kernel selection and interpretability are some weaknesses of SVM [15], [16].

Interpretability means representation of acquired knowledge in a readable form. In Decision trees, it is good. But not in KNN, SVM.

2. Bayesian Networks (BN)

Bayesian (also called Belief) Networks (BN) is a fundamentally important DM technique consisting of two parts, the directed acyclic graph G, with nodes (attributes) and arcs (direct dependencies) and the conditional probability tables for each node [16].

Bayes classifier achieves the optimal result by applying the probability theory. BN represent events and causal relationships between them as conditional probabilities involving random variables. Given the values of a subset of these variables (evidence variables), BN computes the probabilities of another subset of variables (query variables)
However, the Bayesian approaches cannot cross the need of the probability estimation from the training dataset. It is noticeable that in some situations, where the decision is clearly based on certain criteria, or the dataset has high degree of randomality, the Bayesian approaches will not be a good choice.

3. Decision Tree(DT)

Decision tree is a classifier that constructs tree structure with nodes and arcs. Root and internal nodes are labeled with question. Arc represents the answer to the associated question. Each leaf node indicates a prediction of a solution to the problem / value of target variable. Decision tree predicts information in the form of rules which are if-then-else expressions. This outcome explains the decisions that lead to the prediction[16].

4. C4.5

C4.5 is a classifier in the form of decision tree that takes set of data and predicts which class the new data belongs to, based on labeled or training data that are classified already[15].

For a given dataset of patients whose attributes are age, pulse, blood pressure, VO_2max, family history, etc. C4.5 is loaded with these labeled / training data. When a new patient' record with attributes is given as input, C4.5 constructs a decision tree and predicts the class for new patient whether he will get cancer or not, based on their attributes.

This kind of learning is supervised learning, since the training dataset is labeled with classes. Based on this only, C4.5 forms a decision tree and predicts the class of new data.
Difference between C4.5 and decision tree systems

- First, when C4.5 constructing decision tree it uses information gain.
- Second, C4.5 uses a single-pass pruning process to lessen over-fitting. As a result it produces many improvements
- Both continuous and discrete data are used in C4.5. It specifies ranges or thresholds for continuous data thus turning continuous data into discrete data.
- Finally, C4.5 handles incomplete data in its own ways.

5. K-Nearest Neighbors (KNN)

KNN is a classifier. It is a lazy learner. During training process it just stores the training data rather than doing something. When new unlabelled data is given as input, then only it does classification. C4.5, SVM are eager leaner who builds classification model during training. This is supervised learning, since kNN is provided a labeled training dataset[16]. C4.5 builds a decision tree classification model during training. SVM builds a hyper plane classification model during training. First KNN looks at k-nearest neighbors. Second, KNN classifies the data based on idea of the neighbors' classes.

6. k-means

It is a popular cluster analysis that creates *k* groups/ clusters from a set of objects such that the members of a group are more similar. It is unsupervised learning as it "learns" the clusters on its own without any information about which cluster an observation belongs to.
Steps in K-means algorithm

1. K-means picks points in multi-dimensional space to represent each of the k clusters. These are called centroids.
2. Every object will be closest to 1 of these k centroids. They hopefully won't all be closest to the same one, so they'll form a cluster around their nearest centroid.
3. What we have are k clusters, and each object is now a member of a cluster.
4. k-means then finds the center for each of the k clusters based on its cluster members.
5. This center becomes the new centroid for the cluster.
6. Since the centroid is in a different place now, object might now be closer to other centroids. In other words, they may change cluster membership.
7. Steps 2-6 are repeated until the centroids no longer change, and the cluster memberships stabilize. This is called convergence.

Data Mining Tools

There are a number of data mining tools available in the market to fine tune the data mining tasks, using artificial intelligence, machine learning, statistics and other techniques to extract data. Data mining tools can be classified into one of three categories[4]: traditional data mining tools, dashboards,

and text-mining tools. As more number of tools are available, the choice of selecting a suitable tool becomes difficult. Hence having through knowledge of data mining tools is good practice before start the project work. A summary of characteristics of data mining tools is given in table [2].

V- LITERATURE REVIEW

In [17], the authors have studied the techniques that are currently used to analyze Social Media(SM). In this paper the analysis of SM data has been proved to be effective, this is so because of the capacity possessed by data mining in handling unstructured and dynamic data. According to the authors, in future to mine the data generated on SM, research will be carried out on currently used and yet-to-be-explored data mining techniques.

In [13], the author presents the empirical analysis of available data mining techniques to mine social network data. According to author, the future research will focus more on the content mining where lot of human behavior patterns can be identified by analyzing the social network profile pages and also hybrid approach by combining social network analysis (web structure mining) with content mining would be more useful.

In [18], the author has studied on the data mining techniques that are currently used to analyze Social Media data. In this paper, analysis has proved that it is unrealistic to expect one system to mine all kinds of data. Hence different kinds of data mining techniques are available in field for different applications. Every data mining algorithm / technique has its own strengths and limitations.

In [19], the authors have studied the results of the survey papers and have provided some useful ideas for controlling the attributes which mostly affect social media. The Data mining techniques provide a better data control facility. The data mining techniques support for discovering the similarities among the patterns which exist among the voluminous data set. From the outcomes and the results produced, the researchers make a new dimension for the researcher to control the uncontrollable data existing in the social Medias and social networks.

In [2], the authors have summarized that social network data analysis, business and management were the most active domains that requiring mining of social media data and the most frequent social media mining techniques are SVM, BN, and DT. Also the authors suggest that the area of social media still calls for more profound research to house a twin-focus method which incorporates accurate.

In [20], the authors have presented a systematical data mining architecture to mine intellectual knowledge from social data. In this research, they have used social networking site facebook as primary data source and collected different attributes like comments, me, wall post and age from facebook as raw data and used advanced data mining approaches to excavate intellectual knowledge and also analyzed their mined knowledge and suggested that Social data mining is an interesting and challenging research to mine intellectual knowledge which can be used in human behavior prediction, decision making, pattern recognition, social mapping, job responsibility distribution and product promoting.

In [21], the authors have presented a survey on uncertainties of support vectors in SVM. In earlier methods, the values of data point / support vectors are known. Suppose if it is uncertain, SVM becomes more complex in classifying objects as well as in non linear kernel selection. Hence they have suggested that more research could be conducted to deal out this uncertainty of data points and selection of non linear kernel.

VI – OBSERVATION AND DISCUSSIONS

In section IV, various algorithms were discussed. From that, the following information were observed:

Support Vector Machine is very accurate, less over fitting and robust to noise. But interpretability is good in Decision trees, not in KNN and SVM. In Bayesian Networks, it cannot cross the need of the probability estimation, from the training dataset. It is noticeable that in some situations, where the decision is clearly based on certain criteria, or the dataset has high degree of randomality, the Bayesian approaches will not be a good choice. Decision tree predicts information in the form of rules which is easily understandable one. Decision Tree does not handle non numeric data and it needs pruning if it is quite large. C4.5 is quite fast and interpretability is good. It handles both continuous and discrete data. It specifies ranges or thresholds for continuous data thus turning continuous data into discrete data. It uses a single-pass pruning process to lessen over-fitting, as a result it produces many improvements. Over fitting and not working well on small training data set are its limitations. KNN can be quite accurate. It is expensive and requires greater storage. Selecting a good distance metric is crucial to kNN's accuracy. K-Means is faster and more efficient especially over large datasets. But it is sensitive to outliers and the initial choice of centroids, It is designed to operate on continuous data. Extra tricks are needed to work on discrete data.

VII - CHALLENGES AND ISSUES

Extracting actionable and valid patterns through mining is not an easy task. It is very complex and creates many issues [9], [22], [23], [24], [26] and they are categorized as issues like (i) related to data, (ii) tools and techniques, (iii) security, (iv) presentation and visualization of data mining results, (v) Knowledge fusion problem, (vi) institutional commitment and funding

❖ *Related to Data*

1. Poor data quality

Data quality is affected by noisy data, dirty data, missing values, inexact or incorrect values, inadequate data size and poor representation in data sampling. Data cleaning methods and data analysis methods are required to handle noise.

2. Integrating conflicting or redundant data from different sources and forms: multimedia files (audio, video and images), geo data, text, social, numeric, etc…
3. Unavailability of data or difficult access to data.
4. Dealing with huge datasets that require distributed approaches
5. Dealing with non-static, unbalanced and cost-sensitive data
6. Mining information from heterogeneous databases and global information systems.

Since databases are fetched from various data sources available on LAN and WAN. These structures can be in organized and semi-organised. Thus, making them streamlined is the hardest challenge.

7. Processing of large, complex and unstructured data into a structured format.

❖ *Tools and techniques*

1. Efficiency and scalability of data mining algorithms is very important to effectively extract the information from huge amount of data in databases.
2. Constant updating of models is required to handle data velocity or new incoming data.
3. High cost of buying and maintaining powerful software, servers and storage hardware that handle large amounts of data.
4. Sheer quantity of output from many data mining methods.
5. Mining different kinds of knowledge in databases cover a wide spectrum of data analysis
6. Interactive mining of knowledge at multiple levels of abstraction focus the search for patterns, providing and refining data mining requests based on returned results.
7. Incorporation of background knowledge and domain knowledge can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.
8. Data mining query languages and ad-hoc data mining.

Relational Query Languages (such as Structured Query Language (SQL)) allow users to create ad hoc queries for data retrieval. In this area, High-level data mining query languages should be developed to describe ad hoc data mining tasks and also that language should be integrated with a database or data warehouse query language and optimized for efficient and flexible data mining.

9. It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining.
10. Specific data mining systems should be constructed for mining specific kinds of data.

Therefore, one may expect to have different data mining systems for different kinds of data.

❖ *Security*

1. Proliferation of security and privacy concerns by individuals, organizations and governments.
2. Protection of data security, integrity, and privacy.

❖ *Presentation and visualization of data mining results*

1. The presentation of discovered knowledge is very important task in data mining process. It should be expressed in visual representations, or other expressive forms like trees, tables, rules, graphs, charts, crosstabs, matrices, or curves. Hence human can easily understand and apply that knowledge.
2. Data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, either because they represent common knowledge or lack novelty.
3. Evaluation of the patterns: If the pattern data research companies will not be influential and acknowledgeable, it may lack its impact. Therefore, wrong interpretation or even, underestimation can be occurred.
4. Many patterns in DM may be the result of random fluctuations; so many such patterns may be useless

❖ Data Mining of medical data requires specific medical knowledge as well as knowledge of Data Mining technology.

❖ Data Mining requires institutional commitment and funding.

❖ Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem.

VIII- FUTURE RESEARCH WORK

In this information era, Data mining is the most wanted one to extract valuable information / pattern / correlation / trend from huge volume of data which is unstructured and in different format like text, image, video, audio, graphics..etc. Managing and analyzing these kind of data is a great challenge to researchers. As a result, many researchers put their focus in this Data Mining area. Extracted information is applied in many fields like education, medicine, agriculture, banking, sales / marketing, social media.. etc to predict the future event or find the value of target variable. Each and every field needs different kind of knowledge and uses different data repository. Data in data warehouses are heterogeneous, unstructured, dynamic and noisy. Data velocity and size are unpredictable. Hence having one system to mine all these kind of data is unrealistic. Specific data mining systems should be constructed for mining specific kinds of data. Therefore, more numbers of data mining algorithms are available for different kinds of data. Every fraction of a second, huge volume of data is entering into Data ocean. Social Media is the one of the important sources that providing this huge volume of data. These data are

heavily affected by 5 V's characteristics and handling this data is a very big challenge. Many data mining algorithms are used to mine social media data. Most frequently used algorithm is SVM. Non linear kernel selection, training timing, storage and uncertainty of data points are major issues of SVM. Handling these limitations of SVM is a new branch of research.

IX - CONCLUSION

This paper provides idea of data mining including evolution of data mining, data mining parameters, data mining Process, Architecture of data mining system, types of data mining system, data mining algorithms and techniques. All these techniques are having their own merits and demerits. It focuses on social media mining which is the core of this paper. This paper discusses the most frequently used social media mining techniques such as SVM, BN and DT. Due to uniqueness of social media data – velocity, size, dynamism, noisy, unstructured, heterogeneous behavior.etc, researchers are invited to do more research on existing and upcoming technologies. Hopefully in future work there will be further explored in data mining algorithms, including their impact and new research issues.

ACKNOWLEDGMENT

I sincerely thank my Guide Professor Dr. J.Gnana Jayanthi for her guidance and support given.

Table 1. Summary of Merits and Limitations of Data Mining Algorithms

S.No	DM Algorithm	Technique used	No.of Papers Implemented	Merits	Limitations
1	Support Vector Machines	Classification	2,15,16, 18, 21,25	Very accurate , Less over fitting, robust to noise.	Binary classifier, In multi-class classification, kernel selection and interpretability are some weaknesses of SVM, Computationally expensive, runs slow.
2	Bayesian Networks	Classification	2,15,16	Missing data entries can be handled successfully, Over-fitting of data is avoidable	Quality and extent of prior knowledge play an important role, Significant computational cost
3	Decision Tree	Classification	2, 15,16,18	Easy to understand, Easy to generate rules	Over fitting, does not handle easily non numeric data, can be quite large – pruning is necessary
4	C4.5	Classification	15,16,25	Quite fast, Output is human readable.	Small variation in data can lead to different decision tree, does not work very well on small training data set, Over fitting
5	K-Nearest Neighbor	Classification	2,15,16,18, 25	Ease of understanding and implementation, depending on the distance metric, KNN can be quite accurate	Computationally expensive Noisy data can throw off kNN classifications. kNN generally requires greater storage requirements than eager classifiers, Selecting a good distance metric is crucial to kNN's accuracy
6	K-Means	Clustering	2,15,16, 18	Faster and more efficient especially over large datasets	Sensitive to outliers and the initial choice of centroids, It is designed to operate on continuous data – extra tricks are needed to work on discrete data

Table 2. General characteristics of data mining tools

Tool	Language	GUI / Command	Purpose	Merits	Limitations
Weka	Java	Both	General data mining, preprocessing, classification, clustering	Easy to use, Supports different file format - ARFF, CSV, C4.5, binary	Poor representation of result, Not suitable for large data set
RapidMiner	Java	GUI	General data mining, data preprocessing, visualization, predictive	No code required, rich library of functionalities and complete package	limited partitioning ability, Requires knowledge of database handling

			analysis,		
Orange	C++, Python, Qt framework	Both	General data mining, pre-processing, classification, modelling, regression, clustering	Easy to learn, Powerful, support visual programming and Python scripting	Limited reporting capabilities, weak in classical statistics
R	C, Fortran, R	Both	Data mining, statistical techniques	used to make statistical and analytical software, ease of use	Memory management and speed are challenges of R
KNIME	Java	GUI	Data mining, data analysis / Text mining	easy to extend and to add plug-in, a powerful tool with GUI	Limited error measurements, poor parameter optimization

REFERENCES

[1] Hemlata Sahu, " *A Brief Overview on Data Mining Survey*" in International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3, PP: 114-121

[2] MohammadNoor Injadat, Fadi Salo, Ali Bou Nassif "*Data Mining Techniques in Social Media: A Survey*", NEUCOM17295, Volume 214, PP:654-670, 2016.

[3] Aakanksha Bhatnagar, Shweta P. Jadye, Madan Mohan Nagar" *Data Mining Techniques & Distinct Applications: A Literature Review*" in International Journal of Engineering Research & Technology (IJERT), Vol. 1, Issue 9, PP:1-3, 2012.

[4] Dinesh Bhardwaj1, Sunil Mahajan2, "*ANALYSIS OF DATA MINING TRENDS, APPLICATIONS, BENEFITS AND ISSUES*", in International Journal of Computer Science and Communication Engineering, Volume 5 issue 1, PP:53-57, 2016.

[5] Dr. Poonam Chaudhary , "*Data Mining System, Functionalities and Applications: A Radical Review*" in International Journal of Innovations in Engineering and Technology (IJIET), Volume 5, Issue 2 ,PP:449-455, 2015

[6] Neelamadhab Padhy1, Dr. Pragnyaban Mishra 2, and Rasmita Panigrahi3, "*The Survey of Data Mining Applications And feature scope*", in International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, Issue.3, 2012.

[7] Smita1, Priti Sharma, " *Use of Data Mining in Various Field: A Survey Paper*", in IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 3, PP 18-21, 2014.

[8] Mrs. Bharati M. Ramageri," *DATA MINING TECHNIQUES AND APPLICATIONS*", in Indian Journal of Computer Science and Engineering, Vol. 1 Issue. 4, PP: 301-305.

[9] Annan Naidu Paidi " *Data Mining: Future Trends and Applications*" in International Journal of Modern Engineering Research (IJMER), Vol.2, Issue.6, PP:4657-4663, 2012.

[10] Umamaheswari. K, S. Niraimathi "*A Study on Student Data Analysis Using Data Mining Techniques*", in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, PP:117-120, 2013.

[11] Nikita Jain, Vishal Srivastava "*DATA MINING TECHNIQUES: A SURVEY PAPER*" in IJRET: International Journal of Research in Engineering and Technology, Volume: 02, Issue: 11, PP:116-119, 2013.

[12] Ranshul Chaudhary1, Prabhdeep Singh2, Rajiv Mahajan3, "*A SURVEY ON DATA MINING TECHNIQUES" i*n International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, PP:5002-5003, 2014.

[13] S.G.S Fernando et.al "*Empirical Analysis of Data Mining Techniques for Social Network Websites*" in COMPUSOFT, An international journal of advanced computer technology, Volume-III, Issue-II PP:582-592, 2014.

[14] M. Vedanayaki*, "*A Study of Data Mining and Social Network Analysis*" in Indian Journal of Science and Technology, Vol 7(S7), PP:185–187, 2014.

[15] Raj Kumar "*Classification Algorithms for Data Mining: A Survey*" in International Journal of Innovations in Engineering and Technology (IJIET) Vol. 1, Issue 2, PP: 7-14, 2012.

[16] S.Neelamegam."*Classification algorithm in Data mining: An Overview*" in *International Journal of P2P Network Trends and Technology (IJPTT), Volume 4, Issue 8*, PP:369 – 374, 2013

[17] Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F, "*A Survey of Data Mining Techniques for Social Media Analysis*" in Journal of Data Mining & Digital Humanities, PP:1-27, 2014.

[18] Thabit Zatari, " *Data Mining in Social Media* " in International Journal of Scientific & Engineering Research, ISSN 2229-5518 Volume 6, Issue 7, PP:152-154, 2015.

[19] Dr.B.Umadevi1, P.Surya2, "*A Review on Various Data Mining Techniques in Social Media*", in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 4, PP: 8082-8086, 2017.

[20] Rahman, M. M, "*Mining Social Data to Extract Intellectual Knowledge*", in International Journal of Intelligent Systems and Applications(IJISA), vol.4, no.10, PP:15-24, 2012.

[21] XimingWang · Panos M. Pardalos," *A Survey of Support Vector Machines with Uncertainties*", © Springer-Verlag Berlin Heidelberg 2015, Ann. Data. Sci. (2014) 1(3–4) PP:293–309, 2014

[22] MISS. NAZNEENTARANNUM S. H. RIZVI, "*A SYSTEMATIC OVERVIEW ON DATA MINING: CONCEPTS AND TECHNIQUES*" in International Journal of Research in Computer & Information Technology (IJRCIT), Vol. 1, Special Issue 1, PP:136-139, 2016.

[23] Anmol Kumar1, Amit Kumar Tyagi2, Surendra Kumar Tyagi3, "*Data Mining: Various Issues and Challenges for Future :A Short*

discussion on Data Mining issues for future work", in International Journal of Emerging Technology and Advanced Engineering, Volume 4, Special Issue 1, PP:1-8, 2014.

[24] Dipti Verma and Rakesh Nashine," *Data Mining: Next Generation Challenges and Future Directions*" in International Journal of Modeling and Optimization, Vol. 2, No. 5, PP: 603-608, 2012.

[25] Sagar S. Nikam, " *A Comparative Study of Classification Techniques in Data Mining Algorithms*" in ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY, Vol. 8, No. (1): PP: 13-19, 2015

[26] B.R. Patel, "*Comparative analysis of classification algorithm in EDM for improving student performance*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.10, pp.171-175, 2017.

[27]. Nesma Settouti, Mohammed El Amine Bechar and Mohammed Amine Chikh, "*Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classification Task*",in *International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 4, No.1,* PP*:46-51, 2016*

[28] GemaBello-Orgaza,JasonJ.Jungb,∗,DavidCamacho, "*Social bigdata: Recent achievements and new challenges*", http://dx.doi.org/10.1016/ j.inffus.2015.08.005 1566-2535/© 2015 Elsevier,

[29] Shweta Verma, Vivek Badhe, *" Survey on Big Data and Mining Algorithm"*, IN IJSRSET, , Volume 2 | Issue 2 | , PP: 1338-1344, 2016.

[30] Dr.M.Chidambaram, R.Umasundari, "*A Survey on Feature Selection in Data Mining*", in International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Volume-4, Issue-1, PP: 13 -14, 2016

[31] Sunny Sharma, "*A Study on Data Mining Horizons*", in International Journal of Recent Trends in Engineering & Research (IJRTER), Volume 02, Issue 04; PP: 322-326, 2016.

[32] VAISHALI SARATHY, 2SRINIDHI.S, 3KARTHIKA.S, "*SENTIMENT ANALYSIS USING BIG DATA FROM SOCIALMEDIA*", in Proceedings of 23rd IRF International Conference, PP: 40 -45, 2015, Chennai, India.

[33] Parmeet Kaur, " *AN OVERVIEW OF DATA MINING TOOLS*", in International Journal of Engineering Applied Sciences and Technology, Vol. 1, Issue 6, PP: 41-46, 2016.

[34] G Nandi1, A Das1 & 2, " *ONLINE SOCIAL NETWORK MINING: CURRENT TRENDS AND RESEARCH*" in IJRET: International Journal of Research in Engineering and Technology ,Volume: 03 Issue: 04, PP: 346 – 350, 2014.

[35] Wei Fan "*Mining Big Data: Current Status, and Forecast to the Future*" SIGKDD Explorations Volume 14, Issue 2, PP:1-5

[36] H. K. Chan1, E. Lacka2, R. W. Y. Yee3, M. K. Lim4, "*A Case Study on Mining Social Media Data*", in the Proceedings of the 2014 IEEE IEEM, 978-1-4799-6410-9/14/$31.00 ©2014 IEEE , PP: 593- 596

[37] Dave King JDA, "*Introduction to the Mining and Analyzing Social Media Minitrack*", in the proceedings of the 46th Hawaii International Conference on System Sciences, PP :3108-3110, 2013

[38] Albert Ching-man Au Yeung and Tomoharu Iwata, "*Research on Social Network Mining and Its Future Development*" in Feature Articles: Communication Science Reaches Its 20th Anniversary, NTT Communication Science Laboratories, Soraku-gun, 619-0237 Japan, Vol. 9 No. 11, PP:1-4, 2011.

[39] Vidya Shree S I1, Pooja M R2, *" A Review on Data Extraction using Web Mining Techniques*", in International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 4, PP :194-197, 2016.

[40] Mosley Jr, R. C, "*Social media analytics: Data mining applied to insurance Twitter posts*" in Casualty Actuarial Society E-Forum, Winter, vol 2 (p. 1).2012

[41] David Jensen and Jennifer Neville, "*Data Mining in Social Networks*", Papers of the Symposium on Dynamic Social Network Modeling and Analysis. National Academy of Sciences,Washington, DC: National Academy Press. PP:1-13, 2002.

[42] Shubhie Agarwal, Seema Maitrey, Pankaj Singh Yadav, "*A Comparative Analysis of Data Mining Techniques in Wireless Sensor Network*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.4, pp.126-131, 2016.

[43] David Heckerman, "*Bayesian Networks for Data Mining*" in Data Mining and Knowledge Discovery, Volume 1, Issue 1, 1997. Spring 2016

[44] A. Akay, A. Dragomir, "*A Novel Data-Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin*", IEEE J. Biomed. Heal. INFORMATICSournal Biomed. Heal. Informatics. PP:389–396. 2015

[45] Cong Liao1, Anna Squicciarini1, Christopher Griffin2, Sarah Rajtmajer, "*A hybrid epidemic model for deindividuation and antinormative behavior in online social networks*" - Soc. Netw. Anal. Min. (2016) 6:13 DOI 10.1007/s13278-016-0321-5

[46] Bogdan Batrinca • Philip C. Treleaven, "*Social media analytics: a survey of techniques, tools and platforms*", Published online: 26 July 2014 with open access at Springerlink.com, AI & Soc (2015) 30, PP: 89–116, DOI 10.1007/s00146-014-0549-4

[47] Arif Nurwidyantoro," *Event Detection in Social Media: a Survey*", Published in: ICT for Smart Society (ICISS), 2013 International Conference, Date Added to IEEE Xplore: 03 September 2013, INSPEC Accession Number: 13735772.

[48] 1Ms. Pranjali S. Jadhav, 2Dr. Shirish S. Sane, "*UNDERSTANDING STUDENTS' LEARNING EXPERIENCE BY DATA MINING OF SOCIAL MEDIA*", VOLUME-3, ISSUE-4, PP:23-30, 2016

[49] Daljeet Kaur Ȧ *and Aman Paul Ȧ, " *Performance Analysis of Different Data mining Techniques over Heart Disease dataset*", in International Journal of Current Engineering and Technology, Vol.4, No.1, PP: 220- 224, 2014.

[50] K. Jayasudha, "*AN OVERVIEW OF DATA MINING IN ROAD TRAFFIC AND ACCIDENT ANALYSIS*", in Journal of Computer Applications, Vol – II, No.4, PP:32-37, 2009.

[51] P.Veeramuthu, "*Application of Higher Education System for Predicting Student Using Data mining Techniques*", in International Journal of Innovative Research in Advanced Engineering (IJIRAE) ,Volume 1 Issue 5 , PP: 36-38, 2014.

[52] Mohamed Yassine," *A Framework for Emotion Mining from Text in Online Social Networks*", in the proceedings of 2010 IEEE International Conference on Data Mining Workshops, PP: 1136 - 1142, 2010

[53] Aarti Sharma, Rahul Sharma,Vivek Kr. Sharma,Vishal Shrivatava, " *Application of Data Mining – A Survey Paper*", in (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , PP: 2023-2025, 2014.

[54] R.Adaikalam, " *A Survey on Data Mining Techniques for Analysis of Social Network*" in International Journal of Advance Research in Computer Science and Management Studies, Volume 4, Issue 3, PP :65-70, 2016.

Authors Profile

C.Amali Pushpam, MCA., M.Phil., is currently pursuing Ph.D. and servicing as an Assistant Professor in the Department of Information Technology, Bon Secours College for Women, Thanjavur, affiliated to Bharathidasan University, Tiruchirappalli, India since 2006. During her service, she has organized many international and nationsl conferences, seminars and symposium. Her main research work focuses on Data Mining, Big Data Analytics and Network Security.

J. Gnana Jayanthi, M.C.A., M.Phil., Ph.D., is presently servicing as an Assistant Professor in the Department of Computer Science, Rajah Serfoji Government College, Thanjavur, India. She has published more than 30 research papers in International and National conferences and Technical Journals and are cited in popular refereed publishers, IEEE, ACM and Springer. She is a life member of Computer Society of India (CSI), Member of the World Scientific and Engineering Academy and Society (WSEAS), International Association of Computer Science and Information Technology (IACSIT) and member of International Association of Engineers (IAENG). Her research interests include Distributed DBMS, Big Data Analytics and IoT.

